

MODUL 10

CLUSTERING : K-MEANS

10.1. Tujuan

1. Mahasiswa mampu menggunakan algoritma K-Means.
2. Mahasiswa mampu menerapkan algoritma K-Means dalam kasus nyata.

10.2. Landasan Teori

Clustering merupakan suatu teknik data mining yang membagi-bagikan data ke dalam beberapa kelompok (grup atau *cluster* atau segmen) yang tiap *cluster* dapat ditempati beberapa anggota bersama-sama. Setiap obyek dimasukkan ke grup yang paling mirip dengannya. Ini menyerupai penyusunan binatang dan tumbuhan ke dalam keluarga-keluarga yang para anggotanya mempunyai kemiripan. *Clustering* tidak mensyaratkan pengetahuan sebelumnya dari grup yang dibentuk, juga dari para anggota yang harus mengikutinya.

Algoritma K-Means diperkenalkan oleh J.B. MacQueen pada tahun 1976, merupakan salah satu algoritma *clustering* sangat umum yang mengelompokkan data sesuai dengan karakteristik atau ciri-ciri bersama yang serupa. Grup data ini dinamakan sebagai *cluster*. Data di dalam suatu cluster mempunyai ciri-ciri (atau fitur, karakteristik, atribut, properti) serupa dan tidak serupa dengan data pada *cluster* lain.

K-means merupakan salah satu metode *clustering non hirarki* yang berusaha mempartisi data yang ada ke dalam bentuk satu atau

lebih *cluster*. Metode ini mempartisi data ke dalam *cluster* sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu *cluster* yang sama dan data yang mempunyai karakteristik yang berbeda di kelompokan ke dalam *cluster* yang lain. Secara umum algoritma dasar dari K-Means *clustering* adalah sebagai berikut :

1. Tentukan jumlah *cluster*.
2. Alokasikan data ke dalam *cluster* secara random.
3. Hitung *centroid* / rata-rata dari data yang ada di masing-masing *cluster*.
4. Alokasikan masing-masing data ke *centroid* / rata-rata terdekat.
5. Kembali ke langkah 3, apabila masih ada data yang berpindah *cluster* atau apabila perubahan nilai *centroid*, ada yang di atas nilai *threshold* yang ditentukan atau apabila perubahan nilai pada *objective function* yang digunakan di atas nilai *threshold* yang ditentukan.

Distance space digunakan untuk menghitung jarak antara data dan *centroid*. Adapun persamaan yang dapat digunakan salah satunya yaitu *Euclidean Distance Space*. *Euclidean distance space* sering digunakan dalam perhitungan jarak, hal ini dikarenakan hasil yang diperoleh merupakan jarak terpendek antara dua titik yang diperhitungkan. Adapun persamaannya adalah sebagai berikut :

$$d_{ij} = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$$

Keterangan:

d_{ij} = Jarak objek antara objek i dan j

p = Dimensi data

x_{ik} = Koordinat dari obyek i pada dimensi k

x_{jk} = Koordinat dari obyek j pada dimensi k

10.3. Alat dan Bahan

1. Komputer dengan sistem operasi Windows.
2. Program aplikasi RapidMiner.
3. Modul Praktikum Data Warehousing dan Data Mining.

10.4. Langkah-langkah Praktikum

10.4.1. Algoritma K-Means Menggunakan RapidMiner

Contoh Kasus :

Dalam sebuah kelas terdapat 10 siswa yang telah menempuh ujian mata pelajaran Bahasa Indonesia. Data nilai siswa tersebut akan kita gunakan sebagai dasar pengambilan keputusan untuk mencari kelompok siswa yang akan kita kirimkan ke lomba / olimpiade bidang studi Bahasa Indonesia dan Bahasa Inggris.

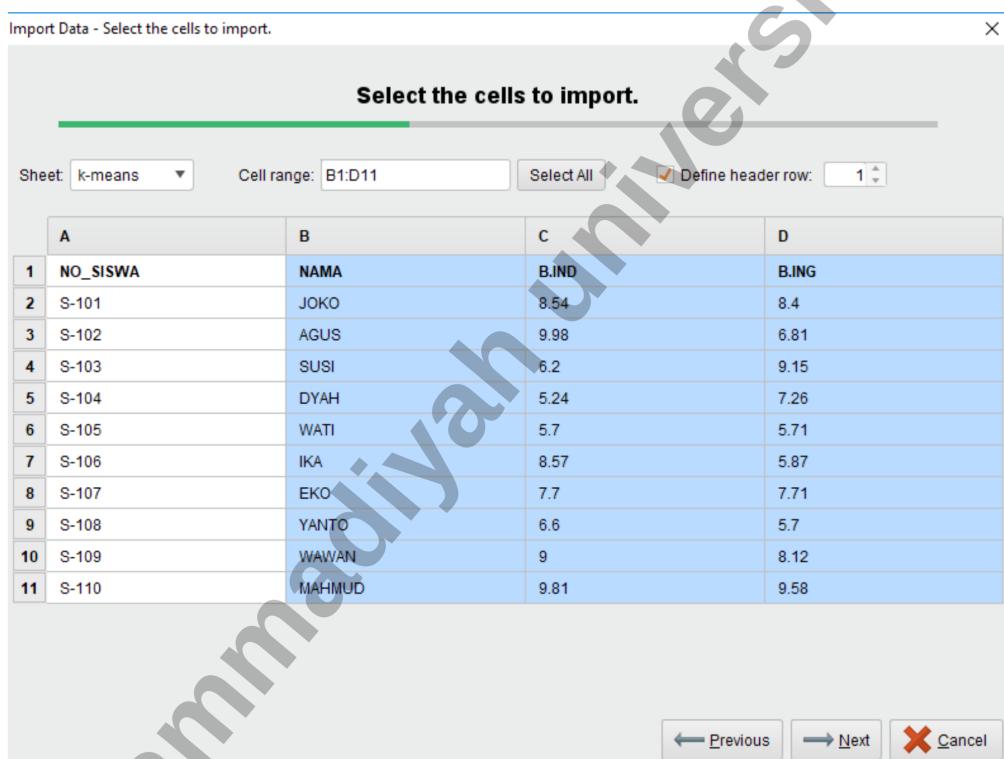
Hipotesis :

Bagaimana mencari kelompok-kelompok siswa dalam bidang studi Bahasa Indonesia dan Bahasa Inggris sesuai dengan nilai ujian yang telah ditempuh oleh siswa.

Berikut tabel data nilai siswa :

NO_SISWA	NAMA	B.IND	B.ING
S-101	JOKO	8,54	8,40
S-102	AGUS	9,98	6,81
S-103	SUSI	6,20	9,15
S-104	DYAH	5,24	7,26
S-105	WATI	5,70	5,71
S-106	IKA	8,57	5,87
S-107	EKO	7,70	7,71
S-108	YANTO	6,60	5,70
S-109	WAWAN	9,00	8,12
S-110	MAHMUD	9,81	9,58

1. Buka Ms. Excel, dan buatlah tabel data nilai ujian siswa tersebut. Simpan dengan nama **Tabel_NilaiUjian.xls** (Format Excel 2003 *.xls).
2. Jalankan aplikasi RapidMiner.
3. Gunakan file **Tabel_NilaiUjian.xls** sebagai data yang akan digunakan dalam proses Clustering. Import file ini ke dalam repositori seperti pada Modul 8 Kegiatan 8.4.2. Pada praktikum ini kita hanya akan menggunakan 3 kolom (nama siswa, nilai bahasa indonesia, nilai bahasa inggris).

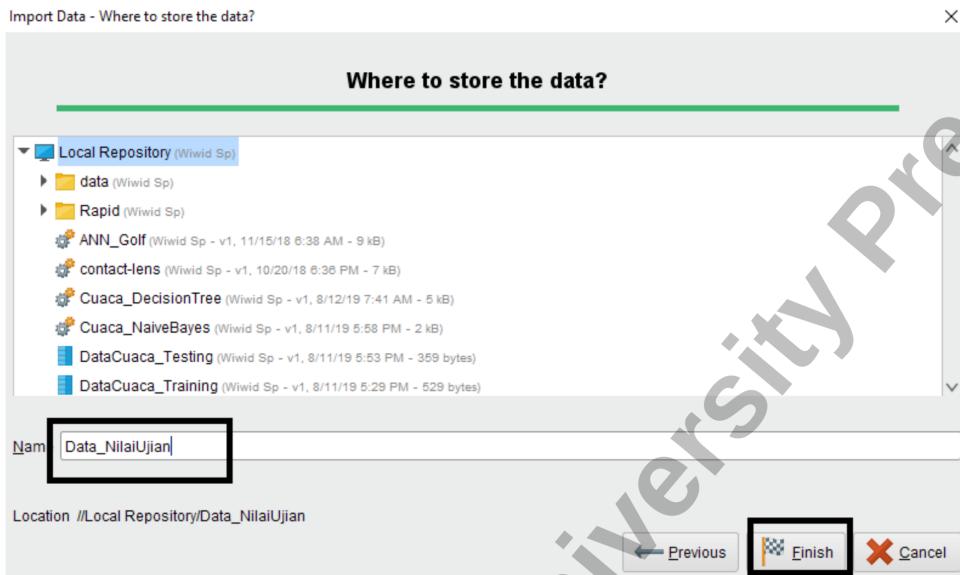


4. Ubah kolom **NAMA** menjadi **id**, dengan cara klik **Change role >> id >> OK**.

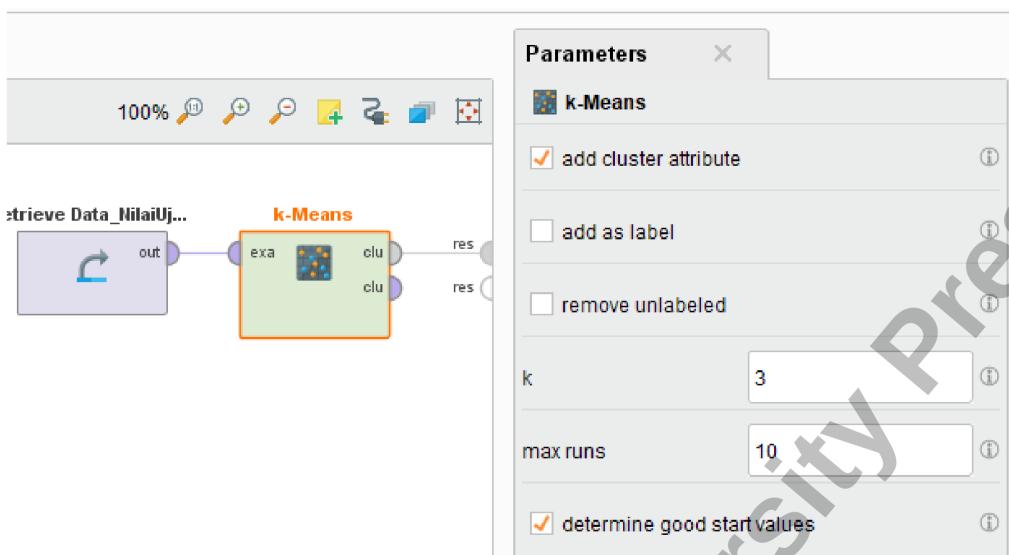
The screenshot shows two instances of the 'Format your columns' dialog box. The top instance is active, displaying a 'Change role' modal. The 'NAMA' column header is circled in black (Step 1). Inside the modal, the 'id' input field is highlighted with a black box (Step 2). The 'OK' button at the bottom right of the modal is also highlighted with a black box (Step 3). The bottom instance of the dialog box shows the updated table where the 'NAMA' column has been renamed to 'id'. The table data is as follows:

	NAMA polynomial id	B.IND real	B.ING real
1	JOKO	8.540	8.400
2	AGUS	9.980	6.810
3	SUSI	6.200	9.150
4	DYAH	5.240	7.260
5	WATI	5.700	5.710
6	IKA	8.570	5.870
7	EKO	7.700	7.710
8	YANTO	6.600	7.700
9	WAWAN	9.000	8.120
10	MAHMUD	9.810	9.580

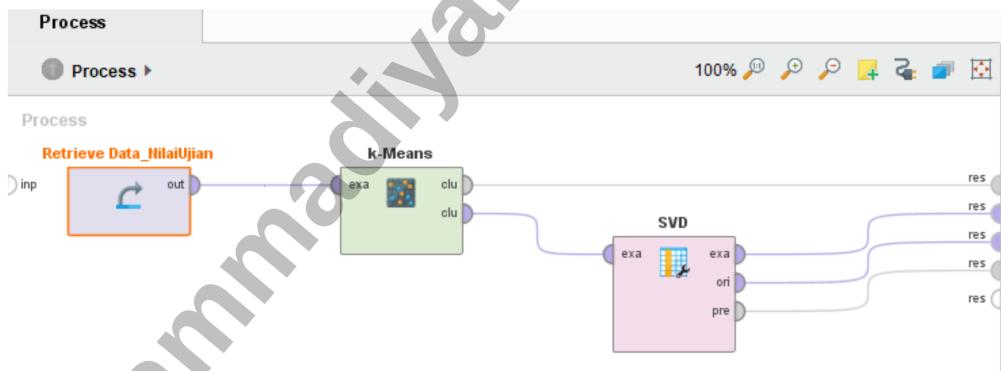
5. Beri nama **Data_NilaiUjian** dan masukkan pada repositories. Kemudian klik **Finish**.



6. Gunakan **Data_NilaiUjian** ini dan masukkan ke dalam area process.
7. Tambahkan operator **k-Means**. Ubah nama operator ini menjadi k-Means. Hubungkan output operator Retrieve ke entry exa operator ini dan output clu (cluster model) dihubungkan ke connector res panel. Ubah nilai parameter k=3 pada operator ini. Angka ini digunakan untuk menentukan jumlah kelompok siswa.



8. Tambahkan pula operator **SVD (Singular Value Decomposition)**. Hubungkan output clu (clustered set) ke-2 operator clustering (k-Means) ke dalam entry exa operator SVD dan 3 port output exa (example set output), ori (original) dan pre (preprocessing model) terhadap connector panel res (result).



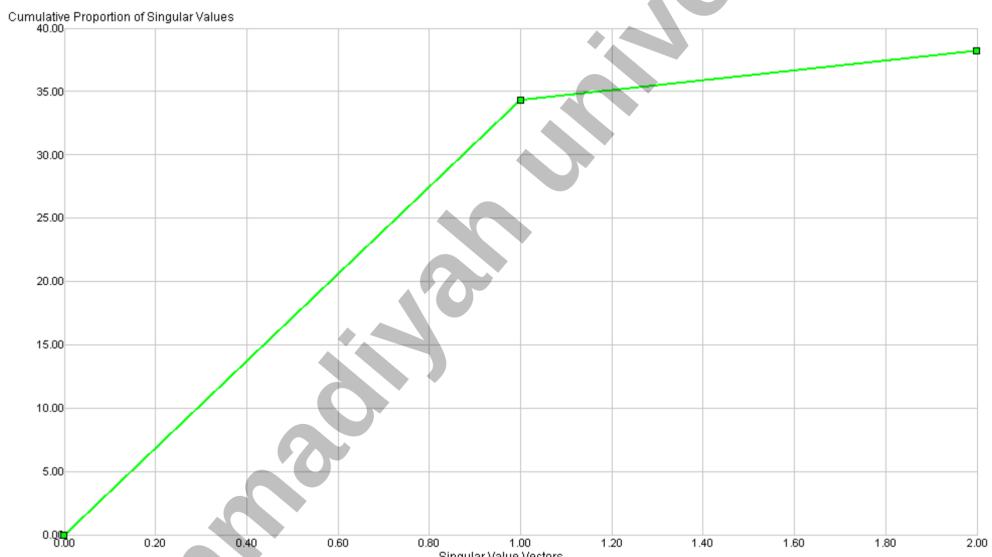
9. Jalankan proses dengan menekan tombol Run (atau menekan tombol F11).

10. Berikut hasil proses Clustering dengan algoritma K-Means.
- SVD (Singular Value Decomposition)
 - Nilai Eigenvalue
 - Nilai Svd vectors
 - Nilai Cumulative variance

Component	Singular Value	Proportion of Singular Values	Cumulative Singular Values	Cumulative Proportion of Sing...
SVD 1	34.340	0.898	34.340	0.898
SVD 2	3.906	0.102	38.246	1.000

Attribute	SVD Vector 1
B.IND	0.723
B.ING	0.690

iii. Nilai Cumulative variance



b) ExampleSet (k-Means)

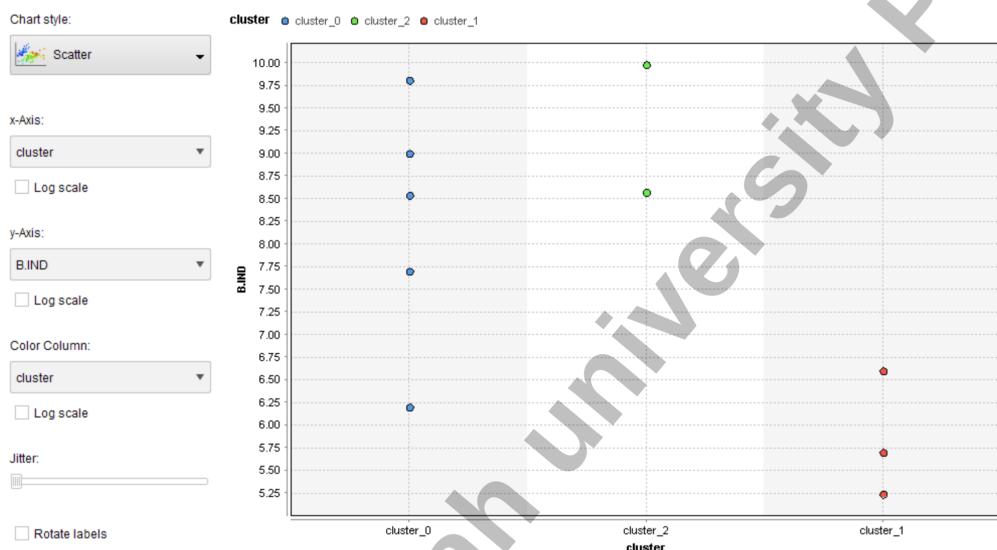
Hasil ini kita lihat dengan mode Plot View menggunakan grafik Scatter untuk menentukan kelompok siswa (cluster) yang dicalonkan untuk maju ke dalam olimpiade mata pelajaran berdasarkan nilai tertinggi ujian.

Ketentuan:

Plotter = Scatter

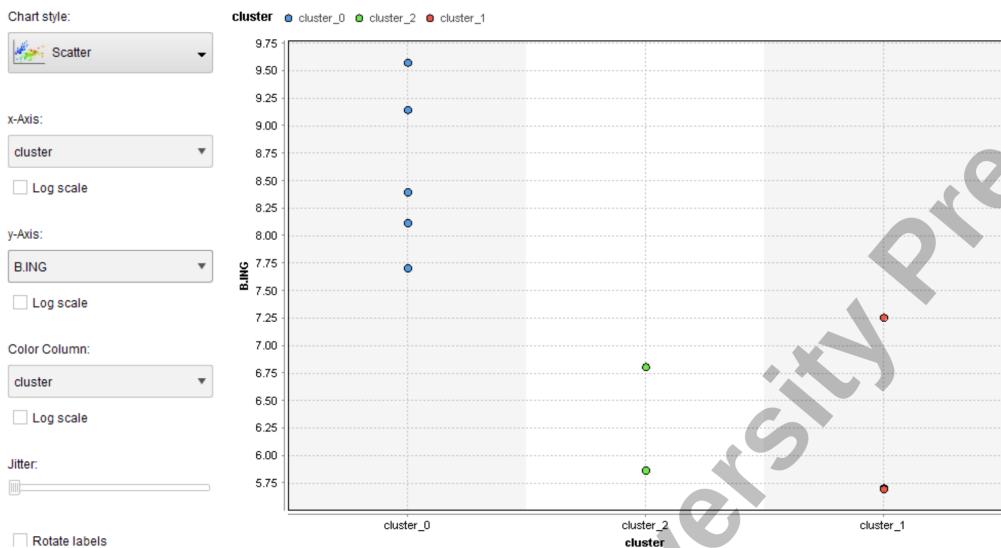
x-Axis = cluster
y-Axis = B.IND, B.ING (diubah-ubah)
Color Column = cluster
Jitter = bisa diubah-ubah untuk melihat distribusi data secara lebih detil.

i. Kelompok Siswa bidang B. Indonesia



Dapat kita lihat bahwa pada cluster_2 merupakan kelompok siswa yang memiliki nilai pelajaran B. Indonesia yang lebih tinggi dibandingkan dengan kelompok cluster_0 maupun cluster_1. Sehingga kelompok cluster_2 yang diajukan untuk lomba olimpiade bidang B. Indonesia.

ii. Kelompok Siswa bidang B. Inggris



Dapat kita lihat bahwa pada cluster_0 merupakan kelompok siswa yang memiliki nilai pelajaran B. Inggris yang lebih tinggi dibandingkan dengan kelompok cluster_1 maupun cluster_2. Sehingga kelompok cluster_0 yang diajukan untuk lomba olimpiade bidang B. Inggris.

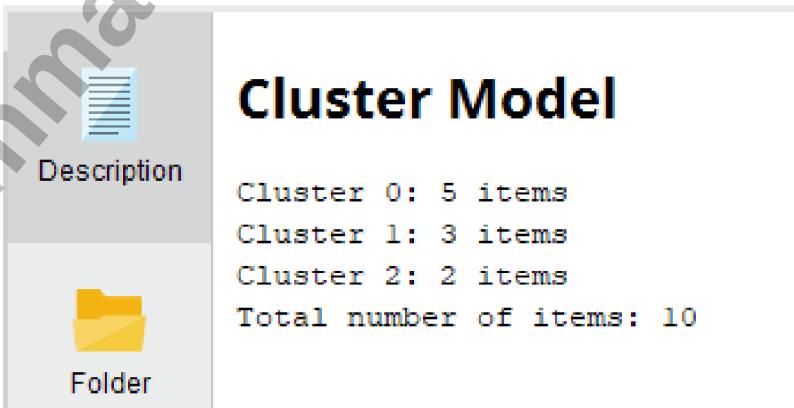
c) ExampleSet (SVD)

Pada hasil ini dilihat secara Data View. Klik pada header kolom *cluster* untuk mengurutkan data berdasarkan *cluster*.

Row No.	NAMA	cluster ↑	svd_1
1	JOKO	cluster_0	0.349
3	SUSI	cluster_0	0.315
7	EKO	cluster_0	0.317
9	WAWAN	cluster_0	0.353
10	MAHMUD	cluster_0	0.399
4	DYAH	cluster_1	0.256
5	WATI	cluster_1	0.235
8	YANTO	cluster_1	0.254
2	AGUS	cluster_2	0.347
6	IKA	cluster_2	0.299

Berdasarkan tabel ini dapat dilihat pembagian kelompok *cluster* siswa. Pada kolom NAMA menunjukkan nama siswa yang terdapat pada data asli.

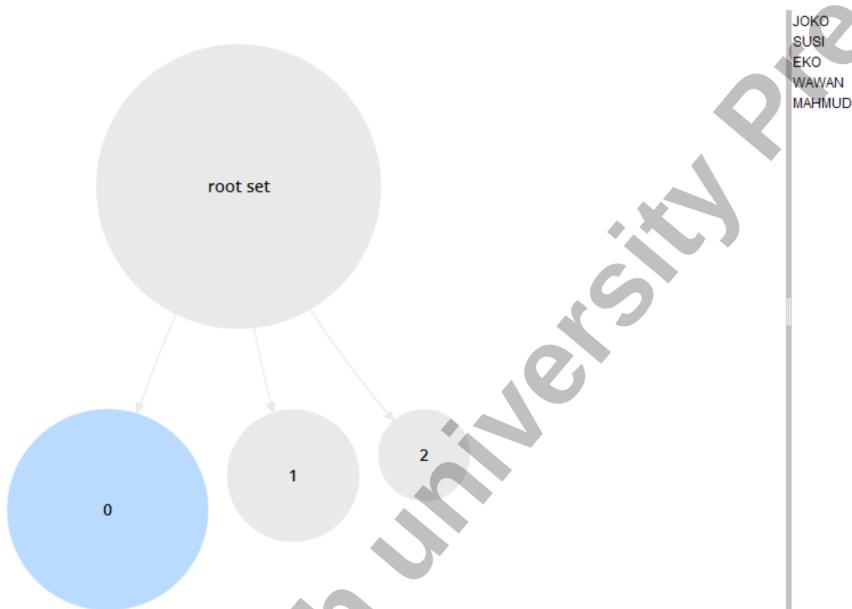
- d) Cluster Model (Clustering)
 - i. Description



Pada cluster model dapat dilihat jumlah data pada masing-masing cluster. Pada cluster 0

memiliki sebanyak 5 siswa, cluster 1 memiliki 3 siswa dan cluster 2 memiliki sebanyak 2 siswa. Dengan total data sebanyak 10 siswa.

ii. Graph



Pada mode ini ditunjukkan bentuk pembagian cluster dengan pola pohon dan cabangnya. Setiap cluster yang terbentuk dapat dipilih (klik) untuk melihat anggota yang terdapat di dalamnya. Pada gambar tersebut dicontohkan pada cluster 0 terdapat 5 anggota siswa yaitu joko, susi, eko, wawan dan mahmud.

10.4.2. Interpretasi Hasil Algoritma K-Means

Berdasarkan hasil kegiatan 10.4.1 dapat disimpulkan pembagian kelompok siswa yang akan diajukan untuk olimpiade Bahasa Indonesia dan Bahasa Inggris adalah sebagai berikut:

CLUSTER	NO_SISWA	NAMA	B.IND	B.ING
0	S-101	JOKO	8,54	8,4
0	S-103	SUSI	6,2	9,15
0	S-107	EKO	7,7	7,71
0	S-109	WAWAN	9	8,12
0	S-110	MAHMUD	9,81	9,58
1	S-104	DYAH	5,24	7,26
1	S-105	WATI	5,7	5,71
1	S-108	YANTO	6,6	5,7
2	S-102	AGUS	9,98	6,81
2	S-106	IKA	8,57	5,87

Pembagian kelompok yang diajukan untuk lomba olimpiade :

1. Cluster_2 yang diajukan untuk lomba olimpiade bidang B. Indonesia.
2. Cluster_0 yang diajukan untuk lomba olimpiade bidang B. Inggris.

10.5. Tugas

Dalam sebuah kelas terdapat 30 siswa yang telah menempuh ujian 4 mata pelajaran, yaitu Bahasa Indonesia, Bahasa Inggris, Matematika, dan IPA seperti dalam Tabel Data Nilai Ujian berikut.

1. Buatlah tabel berikut dengan menggunakan Microsoft Excel!

Tabel Data Nilai Ujian 30 Siswa :

NO_SISWA	NAMA	B.IND	B.ING	MTK	IPA
S-101	JOKO	=5+RAND()*5			
S-102	AGUS				
S-103	SUSI				
S-104	DYAH				
S-105	WATI				
S-106	IKA				
S-107	EKO				
S-108	YANTO				
S-109	WAWAN				
S-110	MAHMUD				
S-111	BUDI				
S-112	SANTI				
S-113	DIAN				
S-114	DANI				
S-115	AHMAD				
S-116	BAYU				
S-117	RISA				
S-118	RANI				
S-119	YANI				
S-120	RATIH				
S-121	INDAH				
S-122	JONO				
S-123	SARAH				
S-124	RAMA				
S-125	BAMBANG				
S-126	HADI				
S-127	NANA				
S-128	FEBRI				
S-129	DENI				
S-130	TONI				

Untuk mengisi daftar nilai dalam tabel, gunakan formula berikut pada salah satu sel. Kemudian bisa di *copy-paste* ke sel yang lain.

$$= 5 + \text{RAND}() * 5$$

(Catatan: setiap mahasiswa pasti akan memiliki data yang berlainan, sehingga hasilnya juga berbeda).

2. Lakukan kembali kegiatan 10.4.1 dan 10.4.2 pada modul 10 ini secara lengkap menggunakan data yang terdapat pada tabel **Tabel Data Nilai Ujian 30 Siswa** tersebut, dengan ketentuan jumlah Cluster = 4. Catat dan tulis semua hasilnya pada lembar jawaban anda, untuk gambar bisa di *copy-paste*.
3. Tulislah masing-masing nama siswa yang terdapat dalam Kelompok Cluster 0, Cluster 1, Cluster 2, dan Cluster 3.